大規模データの俯瞰とターゲットデータの抽出に対する文書 - 単語行列の特異値分解と特異値による重みづけの有効性

平野真理子[†]・小早川　健[††]

東日本大震災ビッグデータワークショップにおいて提供された，震災当日を含めた 1 週間分のツイートのうち，震災対応の初動期間にあたる震災後 72 時間を含む 4 日分のツイッターを解析した．ツイートのクラスタリングによって得られる全体の俯瞰を行ってから目的に応じた分類項目を設定し，その項目に即したツイートを抜き出す抽出器を作成した．一連の作業をよく行うためには，分類項目を設定するために用いられるクラスタリングの性能向上が重要な要素となっている．本研究では，古典的な類義語処理手法である特異値分解をクラスタリングに適用する際に，良く知られている次元圧縮に留まらず，特異値の大きさを特徴量の重みづけの大きさとして活用する手法を提案する．また，クラスタリング結果を人手で修正する作業の容易度を測るための新たな指標を提案し，人手による実作業の効率と比較する実験を行った．その結果，クラスタリングについては，主に作業効率の観点から，特異値による重みづけの有効性と提案する作業指標の妥当性が確認された．分類問題であるターゲットデータ抽出については，学習過程にそもそも重みづけの機構が備わっているにもかかわらず，検出率の向上に若干の効果が見られた．

キーワード：階層クラスタリング，ターゲットデータ抽出，特異値分解 (LSI)

Effect of Singular Value Decomposition and Weighting by Singular Value of Document-Term Matrix, for Large-scale Data Perspective and Targeted Data Extraction

Mariko Hirano[†] and Takeshi S. Kobayakawa[††]

We analyzed tweets broadcasted until four days after the occurrence of the Great East Japan Earthquake, which are provided by the Project 311. After obtaining a general view from tweets clustering, we created a set of targeted extraction categories from them and constructed a tweet extractor tailored to the target. In a sequence of such processes, improvement of the clustering, which is used to discover the target category for extraction, becomes very important. A method is proposed that utilizes the Singular Value as weights for features, while the well-known conventional use of Singular Value Decomposition is limited to reducing its dimension. In addition, we proposed an evaluation criterion for a human-aided clustering task, and conducted experiments to compare these criteria, including commonly-used ones, with the ac-

[†] 株式会社パソナ, PASONA Inc.
[††] NHK 放送技術研究所, NHK Science & Technology Research Laboratories

tual time spent by humans for performing such a task. The experiments show the effectiveness of the proposed weighting method and the competency of our criterion, mainly from the perspective of time efficiency of the task. As for the targeted data-extraction task, which is also a classification problem, some improvement in accuracy is observed although the training process itself involves a weighting mechanism.

Key Words: *Hierarchical Clustering, Target Data Extraction, Singular Value Decomposition (LSI)*

1　はじめに

震災時にツイッターではどのようなことがつぶやかれるのか，どのように用いられるのか，また震災時にツイッターはどのように役立つ可能性があるのか．震災当日から 1 週間分で 1.7 億にのぼるツイートに対し，短時間で概観を把握し，今後の震災に活用するためにはどうすればよいかを考えた．全体像を得た上で，将来震災が発生した際に，ツイッターなどの SNS を利用し，いち早く災害の状況把握を行うための，情報（を含むツイート）抽出器を作成することを最終目標とし，その方法を探った．この最終目標に至るまでの流れと，各局面における課題および採用した解決策を図 1 に示した．図 1 に課題として箇条書きしたものは，そのまま第 3 章以降の節見出しとなっている．

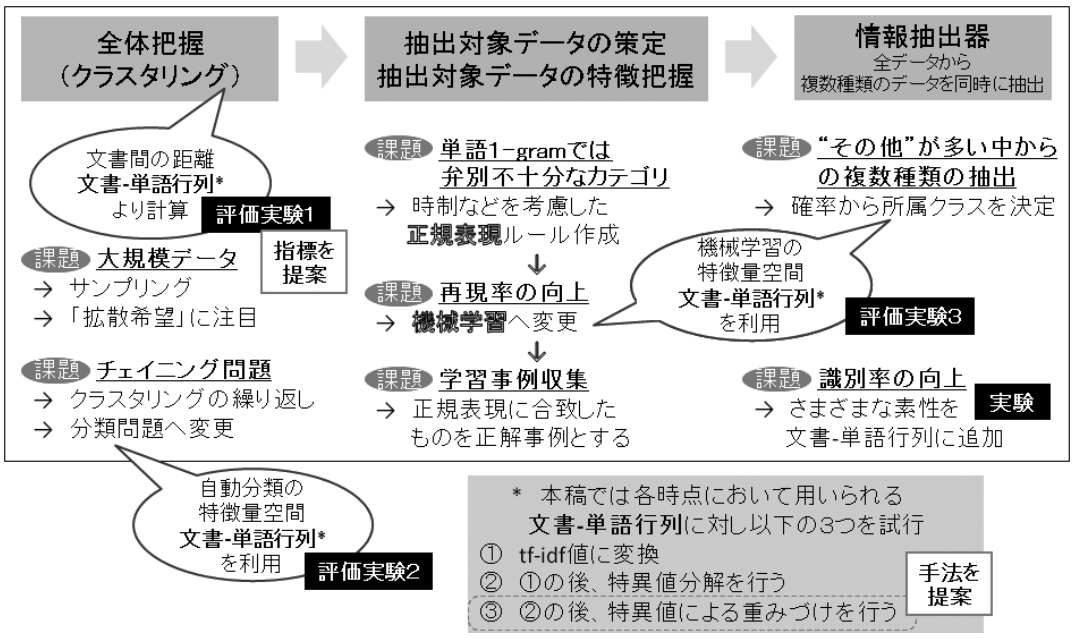

図 1　情報抽出器作成までの流れ

信号処理や統計学の分野において多用される特異値分解は，例えばベクトルで表現される空間を寄与度の高い軸に回転する数学的な処理であり，値の大きな特異値に対応する軸を選択的に用いる方法は，次元圧縮の一手法としてよく知られている．機械学習において，教師データから特徴量の重みを学習することが可能な場合には，その学習によって重みの最適値が求められるが，教師なしのクラスタリングではこの学習過程が存在しないため，特徴量の重みづけに他の方法が必要となることが予想される．筆者らは，本研究の過程に現れるクラスタリングと分類において，古典的な類義語処理および次元圧縮のひとつとしての文書 - 単語行列の特異値分解に加え，特異値の大きさを，特徴量に対する重みとして積極的に用いることを試した．

現実のデータに対し，現象の分析や，知見を得るに耐えるクラスタリングを行うには，最終的に “確認・修正” という人手の介在を許さざるを得ない．この過程で，従来からのクラスタリング指標であるエントロピーや純度とは別の観点からも，文書 - 単語行列に対して特異値分解や特異値による重みづけをすることに一定の効果があることを筆者らは感じた．クラスタリングに多かれ少なかれ見られるチェイニング現象（3.1.3 節で詳細を述べる）を激しく伴うクラスタリング結果は，人手による確認・修正作業に多大な負担をもたらすのだが，このチェイニング現象は特異値分解に加えて特異値で重みづけを行うことで緩和される傾向にあることがわかったのである．そこで本研究では，人手による作業の負担を考慮した作業容易度 (Easiness) というクラスタリング指標を提案し，人手による作業にとって好ましいクラスタリング結果とはどういうものか探究しつつ，文書 - 単語行列の特異値分解と，特異値分解に加えて特異値で重みづけする提案手法の効果，および，従来の指標には表れない要素を数値化した提案指標の妥当性を検証することとする．

以下，第 2 章では，テキストマイニングにおけるクラスタリング，分類，情報抽出の関連研究を述べる．第 3 章では，情報抽出器作成までの手順の詳細を，途中に現れた課題とそれに対する解決策とともに述べる．第 4 章ではクラスタリングの新しい指標として作業容易度 (Easiness) を提案し，それを用いて，クラスタリングや分類を行う際に，特異値分解あるいは特異値分解に加えて特異値で特徴量の重みづけを行うことの有効性を検証する．第 5 章では，「拡散希望」ツイートの 1%サンプリングを全分類して得られた社会現象としての知見と，情報抽出器の抽出精度を上げるために行った試行の詳細およびそれに対する考察を述べる．

尚，本論文の新規性は，タイトルにあるように「文書 - 単語行列の特異値分解と特異値による重み付けの有効性」を示すことであり，関連する記述は 3.1.3 節および第 4 章で行っている．ただし，東日本大震災ビッグデータワークショップに参加して実際の震災時のツイートを解析したこと，すなわち研究用データセットではなく，事後ではあるが，現実のデータを現実の要請に従って解析したこと，によって得られた知見を残すことも本稿執筆の目的の一つであるため，情報抽出器作成の過程全てを記してある．

2 関連研究

テキストマイニングにおいて，クラスタリング，分類，情報抽出の研究は多数存在する (Berry 2004, 2008).

文書の数学的表現としては，ベクトル空間法 (Vector Space Model) が広く用いられている．多次元空間のベクトルを特徴量に用いるというもので，その歴史は古く (Salton 1975) で提案されている．クラスタリングに用いられる文書 - 単語行列は，その自然な拡張である．

単語の表層文字列を素性として扱うような多次元空間では，個々の単語の出現頻度が低く，疎性 (Sparseness) の問題を引き起こす．この問題に対処するために，類義語処理の研究が多数存在しており，次元圧縮としての LSI 法 (Deerwester 1990)，トピックモデルとしての pLSI 法 (Hofmann 1999)，ベイズ推定を用いた Latent Dirchlet Allocation 法 (Blei 2003) が代表的である．

クラスタリングに特異値分解や主成分分析を用いる場合の心得は，(Kobayashi 2004) に詳しい．

LSI 法の数学的な基礎付けとなる特異値分解は，反復法による数値計算で行われる．大規模な疎行例のための実装には，(Dongarra 1978) や (Anderson 1999) がある．

分類問題に対する機械学習の方法としては，線形判別分析 (Fisher 1936)，サポートベクターマシン (Cortes 1995) が代表的である．本研究では，統計処理言語 R による実装 (Karatzoglou 2004) でのサポートベクターマシンを用いる．抽出器の作成には，多クラス分類器 (Crammer 2000; Karatzoglou 2006) を用いて，2 つの手法を比較する．1 つは，サポートベクターマシンの事後確率を計算する方法 (Platt 2000; Lin 2007; Karatzoglou 2006) で，閾値を超える事後確率を持つクラスが存在する場合に抽出する (Manning 2008)．もう 1 つは，1-クラス分類 (Schölkopf 1999; Tax 1999; Karatzoglou 2006) である．

言語処理学会 2012 年度全国大会では，災害時における言語情報処理というテーマセッションで 11 件の発表があった．そこには，効率的な情報抽出という観点から，(Neubig 2012) や (岡崎 2012) の研究がある．本研究は，クラスタリングによって分類カテゴリの決定をするところから始めるという点で，特定の種別の情報を抽出するこれらの研究とは異なる．

東日本大震災時に SNS が果たした役割については (小林 2011), (立入 2011) が刊行されている． (片瀬 2012), (遠藤 2012) でも指摘されているように，今後の震災時に SNS が取材情報源として担う役割は大きいものと思われる．なお，当時のメディアについては，(片瀬 2012), (稲泉 2012), (遠藤 2012), (福田 2012), (徳田 2011) に詳しい．

3 情報抽出器作成までの手順

情報抽出器を作成するまでには，第 1 章の図 1 にあるように，全体把握，抽出対象データの策定・特徴把握，情報抽出というおおまかに 3 つの段階を経た．以下，それぞれについて詳し

く述べる．

3.1　震災ツイートの全体把握

本研究の最終目標は，所望する情報を含むターゲットツイートを抽出する情報抽出器を作成することであるが，そのためには，そのターゲットツイートの持つ特徴をつかむことが必要である．その作業は，そもそもどのような種類のツイートが存在するかを知り，たとえ望ましいまとまりではなかったとしても，実際に形成されたツイート群を見てみることから始まる．つまり，ターゲットデータの抽出のためには，それに先立ち全体傾向を把握すること，すなわちクラスタリングが非常に重要なのである．

3.1.1　大規模データに対するアプローチ（「拡散希望」ツイート）

近年，SNS というメディアの急速な発展に伴って，そこでの発言の解析に関する研究もにわかに脚光を浴びてきている．書き言葉の解析が従来の言語処理のメインターゲットであったのに対し，話し言葉に近い SNS での発言を解析することは，新たな研究課題を含んでいるからである．今回はそのような言語処理の課題に加え，1.7 億というボリュームゆえの大規模データ処理としての課題も顕在化した．

大規模なデータを扱う場合，特定の観点を定めて，それに特化した分析を行うという方法もあるが，我々は始めから特定の観点に限定せずに分析を行いたかったため，ランダムサンプリングを行って全体を把握することにした．サンプリングを行うことで，核心的な個々のツイートを見逃す可能性もあるが，全体を把握する場合は，出現頻度の多いものからとりかかり，その後細部に踏み込んでいくという過程をたどるため，ランダムサンプリングを行うことが自然，かつ効果的である．また，本来“つぶやき”であるものの中から，震災時の状況把握に意味のあるツイートに効率的に接触することを目指し，“拡散させることを目的としている”すなわち“伝える意思が明確である”「拡散希望」ツイートに着目した[1]．実際のところ，キーフレーズ検出[2]を行って検出されたものの中に多数の「拡散希望」ツイートが見られ，震災時に「拡散希望」ツイートが多く出回っていたことも確認されている．

1%ランダムサンプリングを行った上で「拡散希望」ツイートに限定した[3]とはいえ，震災対応初動期間の 72 時間を含む 11〜14 日に限定しても，分析対象のツイートは 3 万件以上あり，全て人手で分類するには 30 人日程度かかることが見積もられた[4]．このため，何らかの自動処

[1] 「拡散希望」に限定しない場合，3 月 11 日〜14 日の総ツイート数は 1.02 億ツイートであり，3 月 11 日〜17 日の全「拡散希望」ツイートは約 385 万ツイートであった

[2] n-gram とそれを構成する各単語 1-gram が個別に出現する頻度の積からなる t 統計量を算出している (Manning 1999)．ここでは，単語 5〜100 gram に対して，t 統計量が 2σ より大きいものをキーフレーズとした．名詞句など特定の品詞列に限定することはしていない．他のキーフレーズ検出アルゴリズムとして (Witten 2005) が有名だが，事前の学習が必要である．

[3] 「拡散希望」に限定したものの中から 1%ランダムサンプリングを行うことと等価である

[4] 筆頭著者の長年にわたる経験では，1 人が 1 日で処理できる自由記述文は 1,000 件程度である

理が必要となったのであるが，この時点ではまだどのような分類項目が存在するかもわからず，加えて時間の経過とともに分類項目が変わっていくことが予想されたため，1 日分ずつ分析対象のツイートのクラスタリングを行うことにした．

3.1.1.1 「拡散希望」ツイートの特徴

「拡散希望」ツイートには次に挙げる 2 つの特徴があり，結果的に「拡散希望」に限定したサンプルツイートは，震災時の膨大なツイートの概観を得るのに非常に有効であった．

特徴 1：基本的には転送を利用して拡散させるため，元ツイートの完全なコピー（公式リツイート）あるいはコピーにオリジナルのコメントを加えたもの（非公式リツイート）が多い

特徴 2：“拡散させたい＝人々にきちんと伝えたい” という意識で書かれているため，一般的なツイートよりも “書き言葉” 寄りで書かれており，スラングや未知語，単語の省略などが比較的少ない．よって，形態素解析における未知語，形態素区切り誤り，品詞誤りも少ない

特徴 1 に関して，今回はリツイートを予め除外しておくことを敢えて行わなかった．リツイートの大きさも一つの情報であり，一つの作業で量と内容を合わせて概観を得るには前もってリツイートのまとめあげを行わない方が適切であると考えたからである．

特徴 2 に関して，3 月 11 日の地震発生後からランダムサンプリングした，「拡散希望」だけからなるツイート 100 件と「拡散希望」を含まないツイート 100 件を調べたところ，前者では全 6,909 形態素中，区切り誤りが 13 件，品詞誤りが 22 件あり，後者では全 3,673 形態素中，区切り誤りが 27 件，品詞誤りが 33 件見つかった．

3.1.2 文書 - 単語行列作成

本研究では一貫して文書 - 単語行列が用いられる．この文書 - 単語行列は，文書（ツイート群）に対し MeCab によって形態素解析を行った後，各文書における単語 1-gram の出現頻度をベクトル空間表現に基づいて作成したものである．続いてこの特徴量ベクトルに対し，キーワードらしさの重みづけに用いられる tf-idf の指標への変換，特異値分解などの処理を行う．これらの文書 - 単語行列に対して行う工夫については，第 4 章においてもう一度説明する．大規模なデータから作成された文書 - 単語行列は，一般的に大規模疎行列になる傾向があるが，本研究では，大規模疎行列に特化したアルゴリズムを用いることなく，一般的な特異値分解のアルゴリズムで事足りた．

本研究では，解析および実験を，統計処理言語 R の標準または一般に入手可能なパッケージに含まれる関数によって行った．表 1 に，本研究で使用した R のパッケージ名，関数，オプションの一覧を，表 2 に，用いた計算環境を示す．

3.1.3　階層クラスタリングのチェイニング現象

クラスタの粒度を任意に設定できる階層型クラスタリングは樹形図（デンドログラム）を用いて視覚的に表現される．根元（図 2 の最上部分）には全てのデータが含まれ，次第に分かれて末端は全てのデータが自分自身のクラスタを形成する．適当な高さ（図 2 破線）で枝刈りをすることで，切断された枝の切断部分より末端に連なるデータがまとまって 1 つのクラスタを形成すると解釈する（図 2 の●または○）．本研究ではユークリッド距離とウォード法を用いてクラスタリングを行った．枝刈りは，クラスタ数が文書数の 1/2 乗になる場所で行うように設計した．ウォード法を用いると比較的チェイニング現象が起きにくいとされているが，著者の経験では，どのような距離関数やクラスタの組み上げ法を採用しても，多かれ少なかれチェイニング現象に遭遇することとなる．チェイニング現象とは，根元から見て，その後更に分か

表 1　統計処理言語 R 使用パッケージ名，関数，オプション等一覧

手順	パッケージ	関数	オプション等
特異値分解	標準（base）	svd	
クラスタリング	標準（stats）	hclust	ユークリッド距離， ウォード法（組み上げ法）
機械学習	kernlab (ver0.9-15)	ksvm （サポートベクターマシン）	rbfdot（ガウシアンカーネル） type = “spoc-svc”（多クラス分類）

表 2　計算環境

CPU	Intel Core i7 960 3.2 GHz
RAM	24GB
OS	Windows 7 (64bit)
アプリケーション	R version 2.15.2 (2012-10-26)

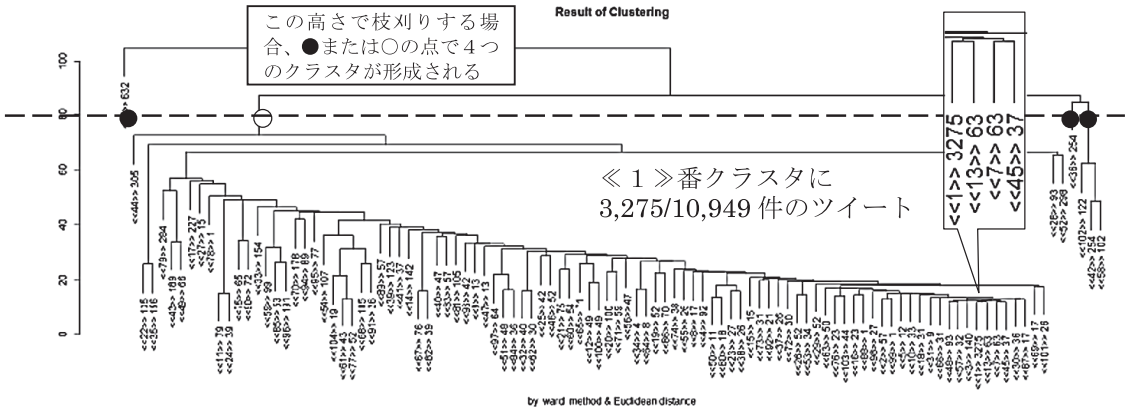

図 2　階層クラスタリングのデンドログラム（樹形図）とチェイニング現象

れることの無い比較的小さいクラスタが次々と分離していき，枝刈りを行った際に，ボリュームが大きく特徴を見出しにくいクラスタ（図 2 の◯印）が残る現象である．

東日本大震災の「拡散希望」ツイートをクラスタリングして顕著だったのは，分離したクラスタのうち，クラスタ内が同じツイートを元とするリツイート群となっているものが多かったことである[5]．リツイート群は出現単語とその頻度が非常に似ており文書ベクトルの距離が近いため，先に分離してクラスタを形成するためと思われるが，逆にこのリツイート群が取得できたことで，人手による分類の確認・修正を行う際，人が見るべきツイートが減ること，また，あるツイートに対する非公式リツイート[6]を含むリツイートの大きさを把握することが可能となること，という 2 つのメリットがもたらされた．リツイートの多さがチェイニング現象の原因であることも考えられたため，図 2 に示している 3 月 11 日の「拡散希望」ツイート 1% サンプリング 10,494 件のうち全くリツイートを含まない（ツイート本文中に文字列 “RT” を含まない）ツイート 667 件に対し，同じ手順で階層クラスタリングを行った．その結果，最も大きなクラスタに 542 ツイート（全体の 81%）が集まるという同様の現象が認められた．リツイートを除かない場合は 10,494 件中 3.275 件 (31%) が最大クラスタに集まっている．

クラスタリングにおいてチェイニングが起きる理由は必ずしも明確ではない (Jain 1999) が，特徴量の設計が不適切で，分類を行う際に弁別能力を持つように文書間の距離を決めらなかった，あるいは階層クラスタリングを行う際の探索アルゴリズムにおいて，得られた解が局所最適解であった，などが原因として考えられる．本研究では，探索アルゴリズムの設計には踏み込まず，統計処理言語 R に用意されている既存のボトムアップ型クラスタリングの関数を利用し，特徴量の設計または用いる距離関数の選択を上手に行い，精度のよいクラスタリング結果を得ることを目指した．

3.1.3.1　階層クラスタリングの繰り返し

「拡散希望」ツイート 1% サンプリングの全分類を目標とし，自立語に限定した単語 1-gram を特徴量とする文書 - 単語行列を作成，クラスタリングを行った．クラスタリングにはベクトル空間表現におけるユークリッド距離を採用し，クラスタ間距離の計算にはウォード法を，クラスタリングアルゴリズムはボトムアップ型（組み上げ法）の階層クラスタリングを採用した．その後，所属文書数が少ない，または文章が短く語彙が少ないクラスタについて，中身を 1 ツイートずつ確認し，ラベルを付与した上で，目視による確認・修正（ラベル付与）が困難なクラスタを集めて再度クラスタリングを行った[7]．これを繰り返せば，全てのツイートにラベルを付

[5] ただし，必ずリツイート群が 1 つのクラスタにおさまることや，1 つのクラスタが 1 つのリツイート群だけで占められることが保証されるわけではない．似ている複数のリツイート群が 1 つのクラスタに入ることもあり，純粋なクラスタリングの目的からすれば，そのようなクラスタの生成こそ理想的である．

[6] 非公式リツイートとは，元のツイートに数文字かそれ以上の追加と削除を加えてツイートしたもので，公式リツイートとは異なり元ツイートにリンクされていないため，機械的なカウントは出来ない（何らかの言語処理が必要）．多くは引用する直前に “RT” の文字を書き加えてある．

[7] 以降，本稿では特に断らない限り，人手が介在する際の作業は筆頭著者が 1 人で行ったことを意味する．

与することが出来るが，繰り返しの手間がかかることはもとより，生成されるクラスタの数が増え続けて全体把握がかえって困難になるのを避けるため，出来上がったクラスタを内容に応じてさらにまとめ上げることが必要となる．また，ボリュームの大きいリツイート群は 1 回目でほぼ出尽くすため，メリットの 1 つであったリツイート群を把握する効果も薄れてくる．そこで，何度もクラスタリングを繰り返すのではなく，2 回クラスタリングを行った後は，それまでに作られたラベルを分類項目として残りをそのいずれかに落とし込むという分類問題に切り替えた．

3.1.3.2　多クラス自動分類の繰り返し

クラスタリングでラベルが付与されたデータを学習データとし，その時点までに作られたラベルを分類項目として，ラベル未定義のデータを対象に，機械学習による多クラス自動分類の識別を行った．興味深いことに，クラスタリング同様半数近くが特定の分類項目に分類されており，そのような項目は所属ツイートが多く，目視で確認・修正作業（ラベル付与作業）を継続することが困難であった．そこで，目視での確認・修正が容易な，ツイート数または語彙が少ない項目に含まれ，確認・修正（ラベル付与）作業が済んだツイートを識別対象から学習データに回し，残りのラベル未定義のツイートに対し，繰り返し機械学習による自動分類を行った．これを 2 回繰り返したところで，各日 9 割の分類が終了した．この際に文書 - 単語行列に対して行った工夫の詳細については第 4 章 4.3 節の評価実験 2 で述べる．分類は，マージン最大化学習であるサポートベクターマシンを採用し，カーネルにはガウシアンカーネルを採用した多クラス分類器を用いた．

3.2　情報抽出器において抽出するべき対象の策定と特徴把握

3.1.3 節においてクラスタリングの確認・修正（ラベル付与）作業を行う過程で，震災時ツイートの分析では “誰が” “誰に” 向かって発言しているか，がより重要な分類軸になることが分かった．もともとソーシャルメディアにおいては，誰もが発信者にも受信者にもなり得，そこで飛び交う情報は，内容も方向も多種多様であるが，特に震災時においては，発信者と受信者の関係性によって情報の担う役割が異なってくるからである．

例えば “被害” に関する話題は，情報の方向を軸に見ると，被災者や被災者から事情を聞いた人が被災地外に向かって被害の状況を説明する “被害実態”，逆に被災地外の人が，テレビが見られない状況にある被災者または被災地周辺に向けて余震や津波の警報を伝える “関連災害予報”，さらに，被災地外から被災者に向けて発せられた，停電時のろうそく使用による二次火災の発生を注意するなどの “二次災害注意喚起” に大別される．同様に，“支援” に関する話題（救出に関するものは別項目）では，発信者と受信者が被災者／非被災者（支援者）のどちらであるかによって “支援を求める声” “支援を申し出る声” “企業や政府に支援を呼びかける声” “支援に関するノウハウを伝える声” などに分けられる．

例えばマスメディアであれば，取材地候補を “被害実態” の中から探し，“支援（物資）を求める声” を人々に伝えるためにツイートを見る．被災者であれば，“支援申し出” の中に，自分が必要としているものが挙がっていないか調べる．これはツイート文中に出現する，個々の被害名称 “地震” “津波” “火災” や物資名の “衣類” “食糧” “粉ミルク” “紙おむつ” といった単語での分類では不十分である．

情報の方向の他に考えられる軸としては，何次の情報であるか（1 次＝本人，2 次＝本人から伝聞，3 次＝間に 1 人介して伝聞）などもあるが，震災時においては情報の方向性がより優先すると筆者らは考えた．そこでクラスタリングによって得られた分類項目（後述になるが第 5 章の表 13 の項目）を整理し，表 3 のように分類項目を再設定した．

今回は特にマスコミが重視する*印の分類項目に注目した．「安否確認」については，マスコミが個々の氏名をツイッターから拾うことはおそらくないものの，連絡不能すなわち通信不能な情報空白地域を特定するために利用することが可能であることから，*印の分類項目とした．

以下に単語による分類から情報の方向性を加味した分類へ分類軸を変更した例を示す．

例 1：「被害」

旧分類項目 地震，津波，火災等その他災害，停電，電話・メール等通信状況……

新分類項目 被害実態，関連災害予報，二次災害注意喚起

（地震，津波，火災，停電，通信状況等は新分類項目の細分項目へ）

表 3　メッセージの方向性を第 1 軸に再設定した分類項目の模式図

大分類	中分類	メッセージの方向		
		被災地→周辺	周辺→被災地	周辺→周辺
被害	*被害実態（小分類：被害種別）	○		
	関連災害予報		○	○
	二次災害への注意喚起		○	
支援	*支援物資要請（小分類：物資）	○		○
	支援申し出		○	
	支援呼びかけ（企業・政府へ）			○
	支援方法・注意点			○
情報（ニュース情報）	*メディア取上げ要望	○		
	情報求める	○		
	*安否確認			○
	情報ソースへ誘導		○	○
	真偽・鮮度の解説・注意			○
メッセージ（クチコミ情報＋応援）	被災生活のノウハウ		○	
	祈り・励まし		○	○

例 2：「支援」

旧分類項目 給水，炊き出し，募金，献血……

新分類項目 支援物資要請，支援申し出，支援呼びかけ，支援方法・注意点

（給水，炊き出し，募金，献血等は新分類項目の細分項目へ）

3.2.1　単語 1-gram のみでは弁別不十分なカテゴリ作成のための素性の追加

クラスタリングは似たもの同士をまとめ上げる機構ではあるが，それがデータ解析に都合のよいまとめ方をしてくれるとは限らない．震災ツイートにおいては，前節で述べたように，単語 1-gram によってまとめただけでは情報活用には不十分であった．筆者らは，単語 1-gram のみによって得られるものとは異なる分離境界を定めての，必要な情報を含むターゲットツイートを抽出する方法を模索した．クラスタリングでは，個々のデータの些末な部分の違いを吸収し，同義語をまとめ上げる必要があるが，このターゲットデータ抽出の段階においては，出現する単語が同じであっても機能や時制や情報の方向性を弁別することが必要となる．そこで，次節に述べるように正規表現を用いて単語 1-gram より長いフレーズの正規表現ルールを書き，目的とする情報を含むツイートを得ることを試みた[8]．

3.2.2　正規表現ルールによるターゲットツイート抽出

クラスタリングによってある程度まとまったツイート群を見渡し，分類項目ごとに単語 1-gram を含む特徴的なフレーズを見出し，人手による正規表現ルールを作成した．先に述べたように，震災時のツイートでは情報の方向を考えてツイートを抽出することがキーポイントになってくるのだが，発信者や受信者が具体的に明記されているツイートは少ない．そこで，情報の方向を暗示する部分（機能表現，時制，共起語）をルールに書き加えることで，収集したいツイートのみが集まるよう工夫した．また今回の災害に限定されないよう，固有名詞や物資の名前等個別具体的な名詞はルールに書き込まないようにした．

例 1：“**火災**が起きています”（被害状況リポート：被災地から周辺へ）

“**火災**が起きないようにブレーカーを落としてから非難を”

（二次災害への注意喚起：周辺から被災地へ）

例 2：“**粉ミルク**が足りません”（支援物資要請：被災地から周辺へ）

“**衣類**を被災地に送るように企業を動かそう”（支援呼びかけ：周辺から周辺へ）

※ ルール化および抽出したのは実線部分のみ．点線部分は特にルール化も抽出も行ってはいないが負例として掲載．

表 4 は，「拡散希望」に限定しない全ツイートからランダムサンプリングで抽出した 3 月 11

[8] 単語 1-gram 以外の新たな素性（例えば，隣接や共起の 2-gram）の投入を試すよりも先に，出現する単語 1-gram に多少の情報を付加して弁別することを試みた．他の素性に関しては 5.2 節に詳述してある．

表 4　正規表現によるターゲットツイート抽出率

データ	設定した抽出目標項目と該当ツィート数（人手で確認）	再現率	適合率	F 値
3 月 11 日 （1,000 件）	・被害実態　142 件 ・安否確認　19 件 ・救出要請　11 件	42.53%	79.57%	55.43%
3 月 13 日 （996 件）	・安否確認　12 件 ・支援物資要請　15 件 ・メディア取上げ要望　9 件	58.33%	87.50%	70.00%

日分 1,000 件と 3 月 13 日分 996 件に対し，両日の代表的な（特にマスコミにとって重要な）分類項目について正解付けを行った後，上記正規表現の抽出率（再現性と適合性）を測定したものである．

3.2.3　機械学習の必要性

表 4 の結果をみてわかるように，正規表現ルールを作成する際は，過検出を防ぐため，適合率重視になりがちである．しかし，人間が発見できない潜在的なルールやうまく書き下すことが困難なルールも存在することは十分予想される．また震災時には素早く情報をつかまなければならないことから，抽出結果の適合率が悪いことは望ましくないが，再現率が悪く，情報にたどりつけないことはそれ以上に大きな問題である．そこで，機械学習を行って再現率の向上を図る必要があるという認識に至った．

3.2.4　学習データ収集の工夫

機械学習には相当数の正解事例が必要である．時間制約のある中で，十分な数の正解事例を一から人手で集めるのは非常に困難である．例えばマスコミが強く関心を持つ “メディア取上げ要望” は，その重要さに相反して人手で精査した「拡散希望」ツイート約 3 万件中には 150 件程度しかなく（後述 5.1 節表 13 参照），それだけでは学習データとしてはかなり少ない．そこで，全ツイートに上記正規表現ルールを適用して集めたツイート群を正解事例として，機械学習を行うことを試みた．このようにして集めたツイート群の中には，実際には抽出対象ツイートでないもの（不純物）も含まれているが，不純物が混入することよりも，学習事例を多く集めることを優先した．こうして集められた正解事例からは，当然ルールに書いた特徴が再び学習されることにはなるが，集められた正解事例に共通する特徴の中には，人間が認識しておらず，明示的にルールに書かれていなかったものも存在するであろう．よって結果的に再現率が向上することをも期待した．

3.3　全ツイートからのターゲットツイート抽出

抽出すべきターゲットツイートとその特徴，および抽出するにあたり注意すべき点を把握したところで，抽出元の範囲を「拡散希望」ツイートから全ツイートに広げた[9]．抽出元の範囲を全ツイートに広げた際に，最も問題となったのは，複数存在する抽出目標のどれにも該当しない “その他” ツイートの存在の多さである．また忘れてはならないのは，災害時に役立つシステムであるためには，情報抽出器は常時稼働，リアルタイム（非バッチ処理），無人で運用されることが想定され，複数種類の抽出を同時に行えるようなシステムにしなければならないということである．

3.3.1　“その他” クラスの扱い

震災後 3 日目になると，震災には無関係なツイートの割合も増えてくる．また，震災に関連してはいても，被災者から発せられている情報のみに注目することにすると，ほとんどが標的外ツイートとなり，“その他” クラスが存在しない一般的な機械学習による分類は困難になる．このタスクは，分類と言うよりはむしろ “その他” の中から目的の情報を抜き出す “抽出” のイメージに近くなる．

初め筆者らは，“その他” をホワイトノイズ的に扱うことを試み，全く脈絡のないツイート群を “その他” クラスの学習データとして与えた．しかし人間には特徴が見出せなくても，機械的に学習される特徴が存在し，それに近い特徴を持つツイートが集められたため，この方法は失敗に終わった．

次に，注目する集合に属するかそうでないかを判定する one-class 分類を複数組み合わせることで複数のターゲットツイート群を同時に抽出することを考え，統計処理言語 R の kernlab パッケージの中のサポートベクターマシン関数 ksvm に用意された type = “one-svc” オプションで実験した．実験対象データは 3.2.2 節で行った実験のうち 3 月 13 日分（全ツイートからのランダムサンプリング 996 件）である．type = “one-svc” オプションでは，ある一つの集合に所属するかどうかが判定されるため，複数の集合のうち，ある一つの集合（仮にクラス A とする）にのみ属し，他の集合には全て “属さない” という結果が得られた場合のみ，そのツイートがクラス A に所属する，という方針で実験を行ったところ，再現率 27.8%，適合率 15.9%，F 値で 20.2%となるなど，結果は芳しくなかった．明確な理由は不明ながら，負例を与えることができないことが一因であることが考えられる．

そこで，関数 ksvm のオプション type = “probabilities” を指定し，予測結果を確率値で出力させ，算出された各クラスタに所属する確率が閾値以上であればそれぞれのクラスタに属するとみなし，どのクラスタに対しても閾値以下である場合はどこにも属さないとする，という定

[9] 先行して行われた 3.2.2 節の正規表現による識別実験も全ツイートからの抽出であった．

表 5　ツイート所属クラスの判定例

	出力された各クラスの所属確率（合計 = 1）			最終判定後の所属クラス	
	クラス A	クラス B	クラス C	（閾値 90%の場合）	（閾値 60%の場合）
ツイート 1	0.96	0.03	0.01	A	A
ツイート 2	0.33	0.33	0.34	なし	なし
ツイート 3	0.12	0.64	0.24	なし	B

義のもと，所属するクラスタを判定する方法 (Manning 2008) を採用した．表 5 にその例を示す．閾値は各実験において，90%から 95%まで 1% 刻みで 6 種類計算し，F 値が最良となるものを都度採用した．閾値を高くすると，クラスに所属すると認定されるツイートが少なくなるため再現率が下がり，閾値を低くすると，クラスに所属すると認定されるツイートが増えるため適合率が下がることが定性的に理解され，また閾値の策定自体も研究課題の一つではあるが、本研究では，どのような文書 - 単語行列が最も識別率を上げるかを問題にしており，それぞれの文書 - 単語行列で最もよい識別率を出す閾値を採用することとした結果，いずれの実験においても閾値 95%が採用された．

3.3.2　複数種類同時に行う「ターゲットツイート抽出」の精度を高める試み

情報の方向を考慮しつつ行うターゲットツイート抽出を複数種類同時に行う，というタスクの精度向上のため，文書 - 単語行列に，単語 1-gram 素性以外にも様々な素性を投入してみた．詳細は 5.2 節において考察とともに述べる．

4　文書 - 単語行列の効果的な変換

3.1.3.1 節，3.1.3.2 節および 3.2–3.3 節において，クラスタリング，自動分類および複数同時抽出には全て文書 - 単語行列が用いられている．自動分類と複数同時抽出はいずれも学習データを用いた機械学習であり，その違いは，選択された分類項目名を出力するのか，全ての分類項目候補に対してそれぞれの確率値を出力するのか，という点だけある．厳密には，そのことに加え，3.1.3.2 節の自動分類は人手によって正確にラベル付与された文書を学習事例としているのに対し，3.2–3.3 節の複数同時抽出では，正規表現でかき集めたことにより混入した “意味内容は該当しない” 事例（不純物）を含む文書を学習事例としている事実がある．ただし，本稿の主旨はこの正解事例の収集方法を比較することではなく，クラスタリング，自動分類，複数同時抽出の各局面において行われた，文書 - 単語行列の変換処理の有効性を示すことである．そこで，この 3 つの局面における変換処理（第 1 章の図 1 に示した①tf-idf 値に変換，②tf-idf 値

に変換した後，特異値分解を行う，③特異値分解を行った後，特異地で重みづけを行う，の 3 段階の処理．第 3 段階が本研究の提案手法）の評価実験を行った[10]．

4.1　文書 - 単語行列の特異値分解 (LSI) と重みづけ

この節に続く 4.2 節，4.3 節 4.4 節は，それぞれクラスタリング，自動分類，複数同時抽出に対して行われた評価実験について詳しく述べたものであるが，ここでは全ての評価実験に共通する，文書 - 単語行列に対して行った 3 段階の処理について説明する．

第 1 段階で行ったのは，作成した文書 - 単語行列を tf-idf 値に変換すること，すなわち文書 - 単語行列にキーワードらしさで重みづけをすることである．これにより，特徴を担う素性の影響力が強化され，出現頻度は高くても，ほとんど全ての文書に登場するような，特徴を担わない素性の影響力を弱められる．この tf-idf 値に変換した行列を X とする．

第 2 段階では X に対して特異値分解を行い，意味軸へのマッピングを行う（式 1）．

$$X = U\Sigma V^T \tag{1}$$

ここで U, V は直交行列，Σ は対角行列となる．特異値分解で文書 - 単語行列は左右の特異値ベクトル（直交行列）と寄与度を表す特異値（対角行列：統計処理言語 R の標準パッケージにある svd 関数の出力では，値の大きいものから順に並んでいる）の積となっているため，式変形を行うと，すでに重みづけがなされているようにも見える（式 2）．

$$XV = U\Sigma \tag{2}$$

ただし $U = [u_1, \cdots, u_n]$，$\Sigma = \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n \end{bmatrix}$ $(\sigma_1 > \cdots > \sigma_n)$ である．

しかし，この段階では tf-idf 値をかけた文書 - 単語行列 X に直交行列 V をかけて単語の軸を意味軸へ変換し，同じ概念を持つ異なる単語を統一的に扱えるようにした[11]に過ぎず，文書間の距離も不変であるため，クラスタリングには何ら影響を与えない．この後，XV のある列より右側をカットした行列，すなわち意味軸へ射影した特徴量のうち寄与度が低い部分を除く “次元縮約” を行った行列を用いてクラスタリングを行うことで，効果的なまとめ上げが可能になる．ただし，カットオフを行うことは，寄与度に閾値 σ cutoff を設け，閾値以上の特徴量を採択し，閾値以下の特徴量を棄却することでしかなく，寄与度の大きさに応じた重みづけはなされていない．

[10] 第 1 段階の tf-idf 値への変換は全ての場合において行うこととし，評価実験は特に行っていない．

[11] この効果をもって，LSI は類義語処理を行っているとも解釈される．また，これは文書と意味軸の関係であるが，同じ行列を転置させた方向から見れば，単語と意味軸の関係が得られ，単語間類似度を算出することもできる．

第 3 段階では，特異値分解を行った（= 単語軸から意味軸へ射影した）文書 - 単語行列に，さらに特異値で重みづけを行う．ここで初めて特徴量に重みづけがなされる（式 3, 4）．なお，第 3 段階で特異値分解の後，特異値で重みづけを行う場合は，特異値分解の後すぐにカットオフを行ってから重み付けを行っても，特異値分解の後，重み付けを行ってからカットオフを行っても同じことであるが，後述するようにカットオフ値が重要なパラメタとなるため，本研究では特異値分解の後，重み付けをしてから最後にカットオフを行っている．

4.2 節以降，“特異値で重みづけ” は

$$XV\Sigma = U\Sigma^2 \tag{3}$$

“特異値の 2 乗で重みづけ” は

$$XV\Sigma^2 = U\Sigma^3 \tag{4}$$

とすることに相当する．

カットオフ σ cutoff をどの位置におくか，言い換えると文書 - 単語行列を特異値分解したものについて，寄与度の大きい方から何列目までを残すかで，その後のクラスタリングや分類の精度が大きく変わることが次節以降の評価実験で明らかになる．詳細はそれぞれ 4.2，4.3，4.4 節で述べる．

(Kobayashi 2004) では，数百単語より大きな規模の問題では LSI の適用が困難とある．しかし，本研究に用いた 32 GByte の RAM を搭載した計算機では，数千単語規模の行列に LSI を適用することに困難はなかった．

4.2　クラスタリングにおける文書 - 単語行列の変換

2 種類の指標によりクラスタリング結果の評価を行う．1 つは従来から用いられているエントロピーと純度によるものである．これらの指標は，計算機による処理のみを行うことを前提としており人手の介在を伴う作業の場合には必ずしも適切な指標ではないと筆者らは感じた．そこで，人手の負担を考慮した評価指標を導入し，人手が介在する場合の機械の処理について検討した．

4.2.1　チェイニング現象の緩和

文書 - 単語行列に対し，特異値分解と，それに加えて特異値による重みづけを行うことで表れる最も顕著な変化はチェイニング現象の緩和である．図 3 を第 3 章 3.1.2 節の図 2 と比較すると，全体の形状からも，枝刈をした時の “残り物” クラスタ（所属文書数最大のクラスタ：図 2 では 3,275 ツイート，図 3 では 1,890 ツイート，いずれも《1》番クラスタ）の大きさからも，特異値分解に加えて特異値で重みづけをすることがチェイニング現象の緩和に役立つことがわ

かる．

図 4 は，3 月 11 日の「拡散希望」ツイート 1% サンプリング 10,949 件に対し，特異値分解なし，特異値分解のみ，特異値分解に加えて特異値の 2 乗および 4 乗で重みづけ，の各場合の，所属ツイート数の多い順に並べた上位 20 クラスタに含まれるツイート数を示したものである．

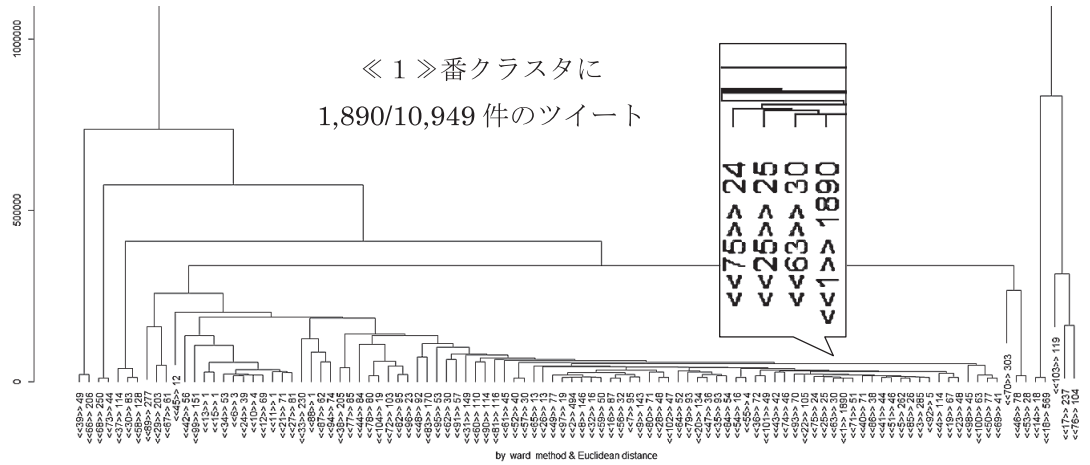

図 3　特異値分解と特異値による重みづけによるチェイニング現象の緩和 (1)

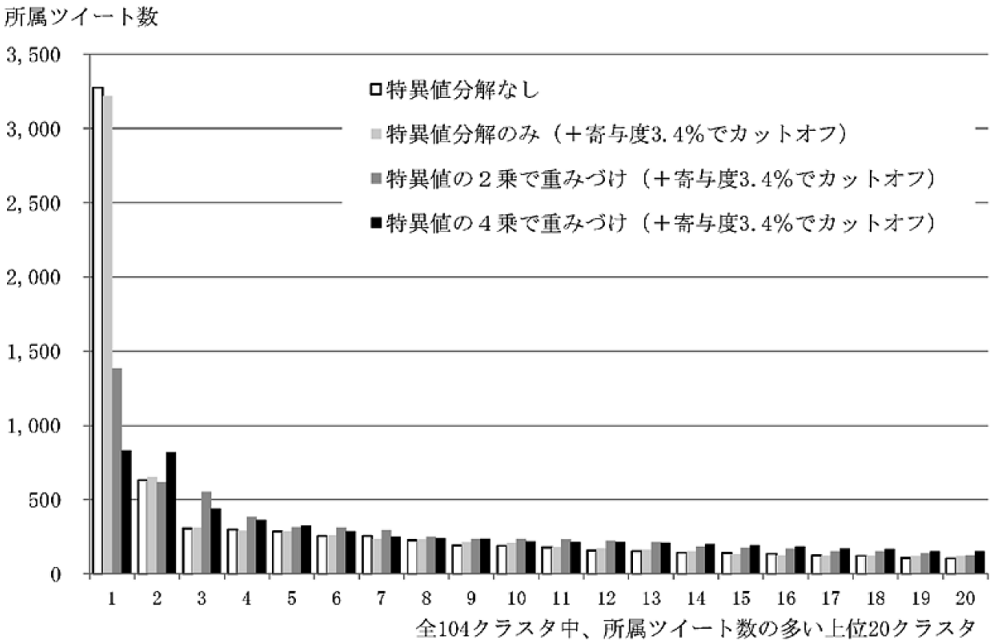

図 4　特異値分解と特異値による重みづけによるチェイニング現象の緩和 (2)

表 6 特異値分解の有無と重みづけの乗数別クラスタリングの評価（3 月 11 日分 10,494 件）

指標	エントロピー	純度	作業容易度	エントロピー	純度	作業容易度
カットオフ σ cutoff	100 列まで（寄与度 4.4%）			160 列まで（寄与度 3.4%）		
特異値分解なし*	0.277	0.720	4.59	0.277	0.720	4.59
特異値分解のみ	**0.251**	**0.767**	4.51	0.275	0.724	4.58
特異値で重みづけ	0.252	**0.767**	4.47	**0.232**	**0.790**	4.47
特異値の 2 乗で重みづけ	0.270	0.764	*4.44*	0.267	0.766	**4.43**
特異値の 3 乗で重みづけ	0.290	0.742	**4.43**	0.296	0.736	4.46
特異値の 4 乗で重みづけ	0.324	0.703	4.48	0.320	0.706	4.49

カットオフ σ cutoff	310 列まで（寄与度 2.4%）			436 列まで（寄与度 2.0%）		
特異値分解なし*	0.277	0.720	4.59	**0.277**	**0.720**	4.59
特異値分解のみ	0.258	0.730	4.58	0.300	0.702	4.64
特異値で重みづけ	**0.229**	**0.790**	4.48	0.304	0.718	4.60
特異値の 2 乗で重みづけ	0.270	0.760	**4.44**	0.326	0.714	*4.57*
特異値の 3 乗で重みづけ	0.298	0.737	*4.45*	0.351	0.685	*4.57*
特異値の 4 乗で重みづけ	0.316	0.704	4.48	0.372	0.666	**4.56**

*特異値分解なし：文書 - 単語行列の全ての行を用いる（カットオフなし）

カットオフは後述 4.2.3 節の表 6 に示す，クラスタリングにおける従来指標（エントロピー，純度）が最良となる値（160 列目，寄与度 3.4%）を用いた．カットオフ前の文書 - 単語行列の列数は 9,013 列であった．特異値分解や，特異値分解に加えて重みづけをすると，重みづけの乗数に応じて各クラスタに所属するツイート数がならされていく様子がわかる．特異値分解を行っただけで重みづけをしていない場合は，特異値分解をしない場合とほとんど差がない．

4.2.2 エントロピーと純度による評価

クラスタリングをする際，文書 - 単語行列に対して行った特異値分解と重みづけに対し，一般的なクラスタリングの評価指標であるエントロピーと純度を算出した．エントロピーおよび純度は，クラスタリング結果と正解のコンフュージョン・マトリクスを作成して比較し，どの程度正解に近い分け方が出来たかを示す指標であり，算出にあたっては，正解が分かっていることが前提となる．結論から言うと，表 6 にあるように，カットオフを適正に定めた場合は，エントロピーと純度から，特異値分解をすること，また特異値分解に加えて重みづけを行うことが有効であることがわかった．

エントロピー：

$$Entropy = \sum_{C_i} \frac{n_{C_i}}{N} entropy(C_i)$$

$$entropy(C_i) = -\sum_{h} P(A_h|C_i) \log P(A_h|C_i)$$

$$P(A_h|C_i) = \frac{x_{ih}}{\sum_j x_{ij}}$$

Entropy：総合エントロピー（各クラスのエントロピーを加重平均）
$entropy(C_i)$: クラス C_i のエントロピー（“正解クラス” の散らばり度）
n_{C_i}: クラス C_i に属する文書の数
N: 全文書数
$P(A_h|C_i)$: クラス C_i と正解クラス A_h の一致度
（クラス C_i に所属する文書のうち正解クラス A_h に分類された文書の割合）
x_{ij}: confusion matrix $(C_i \times A_j)$ の i 行 j 列成分
純度：

$$Purity = \sum_{C_i} \frac{n_{C_i}}{N} Purity(C_i)$$

$$purity(C_i) = \frac{\max(A_h \cap C_i)}{n_{C_i}}$$

Purity: 総合純度（各クラスの純度を加重平均）
$purity(C_i)$: クラス C_i の純度
n_{C_i}: クラス C_i に属する文書の数
N: 全文書数

4.2.3　人手による作業の負担を考慮した評価

Easiness（作業容易度）：
・人手による作業の効率を考慮した新しい指標
・「クラスタリングの確認・修正作業の難易度は，クラスタの質（不純物や語彙の少なさ）だけではなく，クラスタに所属する文書の量にもよる」ことを数値化
・正解データなしでの算出が可能

現在の自然言語処理技術では，精度の良い分類が求められている場合は，少なからず人手の介在（クラスタリング結果の確認・修正すなわちラベル付与作業）が必要である．しかし，人が集中力を途切れさせずにチェック出来るデータ数には限りがあり，せいぜい 100〜200 件程度である．例えば，1,000 件の文書をチェックするのに，1,000 件まとめて一度に作業するよりは，

200 件に小分けされたものを 5 回に分けて作業する方が，心理的負担は少ない．ただし，小分けされていさえすればよいということではなく，クラスタ内は適度に純度が高いことが必要である．クラスタのボリュームの他に，クラスタ内の文章の見た目（出現単語）が似ているかどうかも作業時のストレスに大きく影響する．

自然言語処理における一般的なクラスタリングの精度を表す指標，例えばエントロピーや純度では，上記のような作業効率が考慮されていない．また，エントロピーも純度も，正解付けが行われた後で初めて算出が可能なものであり，人手によるラベル付与作業に入る前に，そのクラスタの出来上がり具合（不純物の多少）を概観することはできない．クラスタの出来上がり具合を知る必要がある理由は，チェイニング現象が激しい場合，所属文書数が多く，従って不純物の多い “残り物” クラスタに対しては，確認・修正のラベル付与作業を行うよりも，そのクラスタだけを取り出してもう一度クラスタリングを行う方が適切であるため，ラベル付与作業を行うかどうか事前に判断しなければならないからである．

以上の経験に基づく考察のもと，作業のしやすさを次の式で与えることとする．

$$Easiness = \sum_{C_i} easiness(C_i)$$

$$easiness(C_i) = \left(\frac{n_{C_i}}{N}\right) \times H_{C_i}$$

$$H_{C_i} = \sum_{W_j} {P_{W_j}}^{(C_i)} \log P_{W_j}^{(C_i)}$$

$$P_{W_j}^{(C_i)} = \frac{{n_{W_j}}^{(C_i)}}{\sum_{W_j} {n_{W_j}}^{C_i}}$$

$Easiness$: 総合作業容易度（各クラスの作業容易度の和）
$easiness(C_i)$: クラス C_i の作業容易度
H_{C_i}: クラス C_i に属する単語のエントロピー（単語で見た場合の乱雑さ）
n_{C_i}: クラス C_i に属する文書の数
N: 全文書数
${P_{W_j}}^{(C_i)}$: クラス C_i 内での単語 W_j の出現頻度
${n_{W_{c_i}}}^{(C_i)}$: クラス C_i に存在する単語 W_j の総出現数

単語 W_i についての和は，クラス C_i に出現するクラスについてのものである．

各クラスごとの作業容易度 $easiness(C_i)$ はそのクラスの所属文書数（を全体の文書数で規格化したもの）と単語エントロピーの積で定義する．値が小さい方が作業が容易である．ここで用いる単語エントロピー H_{Ci} とは，クラスタリングのエントロピー $entropy(Ci)$ とは異なることに注意したい．

この指標の本質は，「クラスタリングの確認・修正作業の難易度は，形成されたクラスタの質

（不純物や語彙の少なさ）だけではなく，クラスタに所属する文書の量にもよる」ことを数値化していることである．またこの指標はエントロピーや純度とは異なり，文書 - 単語行列が作られてさえあれば，クラスタリングが終わった段階で，正解ラベルの付与を行わずとも，そのクラスタまたは全体について算出することができる．中規模のクラスタの確認・修正作業に取り組んではみたものの，実は選り分けが困難なクラスタであったため，半分ほど作業を進めたところで諦め，クラスタリングをやり直さざるを得なくなる，などのような事態を防ぐことが期待される．

表 6 にあるように，次元圧縮のカットオフ σ cutoff が適切な値であれば，特異値分解すること，また特異値分解に加えて特異値で重みづけをすることは，エントロピーや純度の改善に有効であることがわかる．σ cutoff とこれらの指標の関係については今後の研究課題として興味深いところであるが，σ cutoff が適切な値であれば，人手による作業を考慮した提案指標「作業容易度」で見た場合でも，特異値分解に加えて特異値で重みづけをすることが有効であるように見受けられる．4.2.1 節で触れたように，特異値分解後にさらに重みづけを行うこととチェイニング現象の緩和には何らかの相関があることが示唆されているため，この結果は「“チェイニングの度合いが少ない，すなわち枝刈りをした際のクラスタ間の大きさのバランスが取れている” ことが実現されているクラスタリングが，人手による確認・修正作業の作業効率を大きく左右する」ことをも示唆させる．そこで次節では，“人手による確認・修正作業” の評価実験を行った．

4.2.3.1　“人手による作業” の評価 評価実験 1

前節で述べた「クラスタリングに人手が介在する場合は，作業にかかるコストの観点から，“特異値分解” と “特異値分解に加えて特異値で重みづけすること” が有効である」ことを検証するため，3 つのテストセットを用意し，3 人の被験者にクラスタリングの結果を整理してラベルを付与する作業を行ってもらった．各テストセットは，それぞれ 3 月 11 日地震発生以降の「拡散希望」ツイート約 105 万件からランダムサンプリングした 1,000 件のツイート 3 セットで，被験者は各ツイートセットに対して特異値分解をしない文書 - 単語行列でのクラスタリング，特異値分解（意味軸へのマッピング）のみを行った文書 - 単語行列でのクラスタリング，特異値分解を行った上でさらに特異値の 2 乗で重みづけをした文書 - 単語行列でのクラスタリング，のいずれかについて，クラスタリング結果を確認しながらラベルを付与する作業を行った．本研究では，特異値に重みづけをすることの効果を調べることを目的としており，“重みづけ” の代表値として，最良のエントロピーと純度を与えるカットオフ値での，最良の作業容易度を与える “2 乗” を選択した．各被験者はどのテストセットも 1 回ずつ接触し，どのクラスタリング方法も 1 度ずつ経験するようにした．また，各テストセットとクラスタリング方法の組み合わせは $3 \times 3 = 9$ 通りあるが，全ての場合の実験が行われるように実験計画を行った．作業慣れの効果をなるべく減らすため，3 人が経験するクラスタリングの順番はそれぞれ異なっ

ており，作業に伴って現れる疲労の影響を抑えるために，各人実験作業は 1 日に 1 つのみ行うこととした．

各クラスタリング法から見ると，3 種類のテストセットと，3 人の被験者による重複のない 9 種類の実験が行われたことになり，これらを平均することによって，テストセットの内容と被験者の作業能力の差異を吸収させた．また，各被験者が行うクラスタリング法の順が異なるように実験を行い，その結果を平均することで，作業慣れの効果を可能な限り排除した．以上の原則に基づいて割り当てられた 3 人の被験者の実験スケジュールについて表 7 にまとめた．

テストセット 1,000 件にラベルを付与するのにかかった分数と，最初の 60 分でラベルを付与した数を表 8 にまとめた．さらに，それぞれのテストセットに対して 3 者が付与したラベルの一致数と，その割合を表 9 に掲載した．付与すべきラベル数が 35 種類（最初に「拡散希望」ツイートの 1% サンプリングをクラスタリングした際に得られた分類項目の数）とかなり多かったにもかかわらず，平均すると 84% のツイートは 3 人の被験者によって同じラベルが付与されていることがわかる．これにより，必ずしも速度優先で確認・修正作業を行っていたわけではないことが示される．

表 9 から，(1) 特定数のツイートの確認・修正作業にかかる時間で評価しても，(2) 特定の時間内に確認・修正できたツイートの数で評価しても，特異値分解に加えて特異値の 2 乗で重みづけを行った文書 - 単語行列でクラスタリングを行ったものがクラスタリングの確認・修正作

表 7　文書 - 単語行列の特異値分解・重みづけ有無別作業効率に関する評価実験　実験スケジュール

	1 日目	2 日目	3 日目
被験者 A	テストセット A 特異値の 2 乗で重みづけ	テストセット B 特異値分解のみ	テストセット C 特異値分解なし
被験者 B	テストセット B 特異値分解なし	テストセット C 特異値の 2 乗で重みづけ	テストセット A 特異値分解のみ
被験者 C	テストセット C 特異値分解のみ	テストセット A 特異値分解なし	テストセット B 特異値の 2 乗で重みづけ

表 8　文書 - 単語行列の特異値分解・重みづけ有無別作業効率に関する評価実験　実験結果

	最大クラスタの平均ツイート数	1,000 ツイートにラベルを付与するまでの平均時間（分）	最初の 60 分でラベルを付与したツイート数の平均
特異値分解なし	510.00	107.00	708.67
特異値分解のみ	509.00	105.00	711.00
特異値の 2 乗で重みづけ	**397.33**	**88.67**	**808.00**

表 9　各テストセット中，被験者 3 人の付与したラベルの一致数と割合

	3 人のラベルが合致	2 人のラベルが合致	3 人が別のラベルを付与
テストセット A	812 ツイート（81.2%）	156 ツイート（15.6%）	32 ツイート（3.2%）
テストセット B	852 ツイート（85.2%）	134 ツイート（14.4%）	14 ツイート（1.4%）
テストセット C	853 ツイート（85.3%）	123 ツイート（12.3%）	24 ツイート（2.4%）

業を容易にしており，それにはチェイニングの緩和現象が大きく関わっていることがわかる[12]．チェイニング現象の緩和がクラスタリングの確認・修正作業を容易にする一例を挙げると，テストセット B で特異値分解をしない文書 - 単語行列でクラスタリングを行った場合，あるクラスタに分類された津波に関する 19 ツイートは，特異値分解の後重みづけをすると，11 ツイートと 8 ツイートに分離される．8 ツイートは，全て “停電で宮城の人は大津波警報知らないそうです” というツイートのリツイートになっており，よりクラスタリングが細分化されたことになる．一方，特異値分解をしない文書 - 単語行列でクラスタリングを行った場合に生成された “公衆電話が無料になりました！携帯電話使えない方ぜひ利用して！” というツイートのリツイートが集まっていたクラスタに，特異値分解の後重みづけを行った文書 - 単語行列で再度クラスタリングを行うと，“携帯電話よりも公衆電話の方が繋がります” “現在公衆電話が国内通話無料解放中です．回線が優先的に繋がるようになっているので付近の方は公衆電話を利用しましょう” という 2 ツイートがそのクラスタに合流した．これにより，このクラスタは 1 つのリツイートの集合ではなくなったが，ツイートの内容は殆ど同じである．このように，特異値分解や重みづけを行うと，クラスタの分離と統合両方が生じるが，重要なことは，分離・統合を経てもクラスタの内容の均一性が保持されるということである．以上により，今回提案した人手による作業の負担を考慮した評価指標 Easiness が作業効率と同じ傾向を持つことが示され，同時に，正確さを確保しながら（すなわち人手による確認・修正を加えながら）素早く分類を行う必要がある場合には，チェイニング現象を抑えておくことが有効であること，また，特異

[12] (1)，(2) それぞれにおいて，(a) 特異値分解なし，(b) 特異値分解のみ，(c) 特異値分解 + 重みづけの 3 つの手法ごとにまとめた実験結果に対し，平均値の差の検定を行った．具体的には，被験者やテストセットが共通しており，また事前に行った分散の等質性の検定により等分散の仮定が妥当であると判定されたため，“対応のある t 検定（自由度 2）” を行った．(1) においては，(a) と (b) の差の p 値は 0.8164，(a) と (c) の差の p 値は 0.07284，(b) と (c) の差の p 値は 0.2999 であり，(2) においては，(a) と (b) の差の p 値は 0.9691，(a) と (c) の差の p 値は 0.06231，(b) と (c) の差の p 値は 0.3207 であった．(1)，(2) いずれにおいても，有意水準をどこに置くかにかかわらず，(a) と (b) には有意な差があるとは認められない．つまり，特異値分解を行っただけでは確認・修正作業の効率に差は生じない．一方，(a) と (c) は，最も広く用いられる有意水準 5% を採用した場合は有意差なしと判定されてしまうものの，社会科学等で用いられる有意水準 10% を採用した場合は有意差ありと判定される．主観評価実験であることから，社会科学寄りの基準を用いてもよいとするのであれば，有意差があると判断しても良い．また，今回の評価実験においては，被験者 3 人のうち 1 人のみがラベル付与作業に精通しており，他の被験者に対する実験後のヒアリングでは，「35 種類ものラベルを付与することに戸惑ったが，回数を重ねるごとに慣れてきて作業が容易に行えるようになった」という回答も見られた．実験計画法に基き「慣れ」の影響を排除するよう努めてはいたが，予想以上にこの効果が強かったことが，有意差が微妙であったことの最大の原因と考えられる．実験数を増やした上で最初の数回分の実験データは対象外とする，という方法を取る必要があった．

値分解に加えて特異値で重みづけした文書 - 単語行列でクラスタリングすることで，自動生成されるクラスタの質をそれほど損なわずにそのことを実現出来る，ということも示された.

4.3　自動分類における文書 - 単語行列の変換 評価実験 2

クラスタリングを数回行った後，“残り物” クラスタに含まれる文書数がまだ多く，かつ人手で振られるラベルの種類が限定されてきた段階では，クラスタリングの確認・修正作業で整えられたクラスタのラベルを分類項目として，ラベル未定義の文書をそのいずれかに振り分ける自動分類を行うことで，効率的に分析対象のツイートを全分類することが出来る．評価実験 2 では，3 月 11 日分の「拡散希望」ツイートの 1% サンプリングのうち，2 回クラスタリングを行い，2 回自動分類を行ってなおラベルが定義されなかった 1,141 件のツイートに対して分類項目の識別実験を行った（表 10）．分類項目が 35 と多かったためか，既に自動分類を 2 回行った後の残りのツイートに対する分類問題であったからか，全体的に識別率はそれほどよくない．表 10 を見る限り，4.1 節で述べたカットオフ値（特異値分解，重みづけを行った文書 - 単語行列の何列目までを用いるか）を適切に選ばない限り，分類問題に対しては特異値分解を行うことは識別率の改善にはつながらない．また，特異値分解のみと特異値分解に加えて重みづけをすることに，それほど差はない.

4.4　複数同時抽出における文書 - 単語行列の変換 評価実験 3

情報抽出器において複数同時抽出を行う際にも，文書 - 単語行列に対し特異値分解と特異値による重みづけが有効かどうかを調べた.

3.2.2 節で用いたのと同じ，全ツイートからランダムにサンプリングして正解付けを行った 3 月 11 日分 1,000 件と 3 月 13 日分 996 件に対し，両日の代表的な（特にマスコミにとって重要な）分類すべき項目について，複数種類の同時ターゲットツイート抽出実験を行った．クラスタリングと同様，tf-idf 値に変換した文書 - 単語行列，それに加え特異値分解を行い，意味軸へのマッピングを行ったもの，さらに特異値で重み付けを加えたもので抽出率を比較した．特異値分解，および特異値分解の後重みづけを行った文書 - 単語行列に対しては，寄与度 2% 以下

表 10　文書 - 単語行列の特異値分解・重みづけ有無別　35 項目の自動分類の識別率

文書 - 単語行列	カットオフ			
	100 列まで（寄与度 4.4%）	160 列まで（寄与度 3.4%）	310 列まで（寄与度 2.4%）	436 列まで（寄与度 2.0%）
特異値分解なし*	51.2%	51.2%	51.2%	51.2%
特異値分解のみ	49.0%	**52.6**%	43.6%	37.6%
特異値の 2 乗で重みづけ	49.0%	**52.6**%	43.6%	37.4%

*特異値分解なし：文書 - 単語行列の全ての行を用いる（カットオフなし）

表 11　特異値分解の有無と重みづけの乗数別　ターゲットツイートの抽出率（3 月 11 日分 1,000 件）

	所属クラス判定の閾値	再現率	適合率	F 値
特異値分解なし	95%	54.07%	23.79%	33.04%
特異値分解のみ	95%	39.53%	41.72%	40.60%
特異値で重みづけ	95%	38.37%	43.14%	40.62%
特異値の 2 乗で重みづけ	95%	37.21%	44.76%	40.63%
特異値の 3 乗で重みづけ	95%	38.95%	42.68%	40.73%
特異値の 4 乗で重みづけ	95%	38.95%	44.37%	**41.49**%

表 12　特異値分解の有無と重みづけの乗数別　ターゲットツイートの抽出率（3 月 13 日分 996 件）

	所属クラス判定の閾値	再現率	適合率	F 値
特異値分解なし	95%	72.22%	30.59%	42.98%
特異値分解のみ	95%	75.00%	38.57%	50.94%
特異値で重みづけ	95%	72.22%	40.00%	51.49%
特異値の 2 乗で重みづけ	95%	77.78%	42.42%	**54.90**%
特異値の 3 乗で重みづけ	95%	77.78%	40.58%	53.33%
特異値の 4 乗で重みづけ	95%	77.78%	40.58%	53.33%

の列を削除する次元圧縮を行った．単語 1-gram 素性には，自立語のほか，助動詞，副詞，連体詞，接続詞，接頭詞，感動詞を用い，助詞と記号以外のほとんどを用いることとした．クラスタリングとは異なり，情報の方向性を，自立語以外のさまざまな部分から得られる手がかりで弁別する必要性があったからである．素性に単語 1-gram 以外のものを追加した試みの詳細については 5.2 節で述べる．

実験の結果，ターゲットツイートの抽出に対し，特異値分解に加えて重みづけを行うことが有効なことがわかった（表 11，表 12）．第 1 章で述べたように，そもそも学習を行うと，特徴量には最適な重みづけが行われるため，特異値分解に加えて重みづけすることの効果はあったとしても薄れるはずである．しかし，実験の結果から，特異値分解に加えて重みづけをすることの効果がないわけではないことが見てとれる．F 値で見る限り，重みづけの乗数は 2 乗付近がピークになっており，クラスタリングの実験結果と合わせ，特異値による重みづけは 2 乗程度が適当であると考えられる．11 日の方が再現率が低いのは，“被害実態”という多種多様なツイートが所属しうる分類項目に対しての抽出実験であったため，各項目ごとの学習事例数をそろえて抽出を行った今回は，“被害実態”の細分項目 1 つあたりの学習事例数が相対的に少なくなってしまったことや，特異値分解の効果を大きく左右するカットオフの設定が適切ではなかったことが原因と思われる．

なお，表 11，表 12，表 14 はいずれも所属クラス判定の閾値は全て 95% となっているが，これは各文書 - 単語行列に対する実験で，その都度最良の F 値を与える閾値を採択した結果である．

5 情報抽出器作成の過程で得られた知見と残された課題

本研究の最終目的は，災害発生時に役立つ情報抽出器を作成することであったが，それに先立つ全体把握のためのクラスタリングを完遂したところで見えてきたこと，抽出器の抽出精度を挙げる段階で課題として残ったものがあるため，ここに記しておく．

5.1 「拡散希望」ツイートの分類によって得られた社会現象としての知見

「拡散希望」ツイートの 1% サンプリング 11〜14 日分 3 万件の 9 割を分類してみて最も驚いたのは，そこに人間の善性が表れていたことである．「ケガ人の手当ての仕方」「救出を待つ間にするべきこと」「被災生活のサバイバルノウハウ」など，マスメディアによる報道には取上げられることの少ない，口コミ系メディア特有のツイートが数多く存在した．また，マスメディアが取り上げきれない地域の細かい情報をまとめ，ウェブ上に掲載する人が少なからずいる一方，散在するそれらの情報を必要としている人に届けようと，かなりの人が進んで仲介の役割を担ったことがうかがえる．直接的に「生きろ！」と叫ぶ声，被災・非被災にかかわらず，全ての人に「元気を出そう」と励ます声も，情報的な価値と関係無く「拡散希望」の対象となった．「企業に働きかけて，被災地に支援物資を送らせよう」という運動さえ起こっていた．そこには情報収集のツールを使い回す人々の姿よりも，本心から被災者を思いやり，助けようとする人間的な温かさを持った人々の姿のイメージがあった．表 13 は「拡散希望」ツイートの 1% サンプリング 11〜14 日分 3 万件の全分類結果（各日 1 割程度が未分類）である．表 13 の分類項目は，クラスタリングの結果をもとにしており，情報の方向については特に考慮していない．

5.2 災害用情報抽出器における，文書 - 単語行列の素性についての考察

災害発生時における，放送等マスメディアにとって有用な情報収集のためには，情報の方向で抽出目標のツイートを弁別することが大切であることは 3.2 節で特に詳しく述べた．このため，単語 1-gram 素性に加え，①機能表現，②動詞文節（連続する動詞と助動詞はひとまとまりにする），③個別正規表現（分類項目ごとの正規表現集が全体として当たったかどうかではなく，正規表現集の個々の表現に対する合致／非合致），④正規表現で集めたツイート群に含まれる 5-gram 前後のキーフレーズ，⑤隣接単語 2-gram，⑥自立語に限定した共起 2-gram，⑦正規表現ルールに含まれる単語とそれを特異値分解にかけて得られた類義語による限定共起 2-gram，を文書 - 単語行列に追加し，それぞれの効果を調べた（表 14）．

予想では，多義を持つ素性である単語は再現率が高く，逆に時制や意思を表す機能表現や，文脈を形成する共起語などは，意味解釈を限定していく作用があるため適合率が高くなると思われたが，一見するとそのような効果は見られなかった．しかし結果をよく見ると，一つの傾向が見えてくる．

表 13　「拡散希望」ツイート（1%サンプリング）全分類

	4 日間計	3 月 11 日	3 月 12 日	3 月 13 日	3 月 14 日
総計	30,862 件	10,949 件	9,250 件	6,387 件	4,276 件
イベント中止情報	89	40	28	10	11
地震	543	390	110	39	4
津波	694	596	41	6	51
火災，爆発，噴火，その他災害	320	138	113	29	40
原発	473	46	199	174	54
停電（計画停電）	444	45	104	115	180
節電	2,045	145	1,539	267	94
電話—171 含む—（使い方，現況）	1,738	1,412	199	47	80
メール（使い方，現況）	313	268	20	25	
ツイッター（使い方，現況）	593	170	207	116	100
公共交通（主に都市部）	392	315	58	5	14
被害状況レポート	887	72	116	576	123
安否	1,937	165	625	678	469
救出要請	1,450	814	501	85	50
救出待つ人へ	760	511	239	10	
施設情報（開放，営業）	1,402	970	184	120	128
物資情報（無料配布，給水）	690	232	263	150	45
炊き出し	206		122	11	73
買い占め	120	3	4	6	107
募金・チャリティー	519	2	71	242	204
献血	254	1	155	96	2
支援一般（回収場所，注意点等）	369	3	210	102	54
支援申し出	169		107	45	17
支援呼びかけ（企業・政府へ）	464		13	376	75
支援物資求める	732	33	198	237	264
メディア取上げ要望	146	3	1	27	115
情報求める	102	13	13	38	38
情報ソースへの誘導	2,717	585	801	662	669
情報管理（デマ指摘含む）	842	450	288	88	16
義援金詐欺・不正情報	114		31	41	42
被災者へ（避難生活の知恵）	1,439	542	590	214	93
帰宅難民へ	24	24			
火災防止	590	455	115	4	16
手当て法	426	242	92	64	28
道空けよう	668	231	221	108	108
治安の呼びかけ	1,142	493	173	351	125
政府へ	75		1	17	57
マスコミへ	144	6	42	16	80
外国人・障害者へ	763	358	213	171	21
全ての人へ（祈り・応援）	724	2	283	331	108
遺族へ	18			12	6
その他（震災と無関係）	481	33	89	114	245
未分類	2,844	1,141	871	562	270

表 14 素性別抽出分類問題の精度（3 月 13 日分 996 件）

文書 - 単語行列に用いた素性	所属クラス判定の閾値	再現率	適合率	F 値
単語（1-gram）のみ	95%	75.00%	38.03%	50.47%
単語＋機能表現	95%	66.67%	32.43%	43.64%
単語＋動詞文節	95%	66.67%	33.80%	44.86%
単語＋個別正規表現	95%	66.67%	39.34%	49.48%
単語＋キーフレーズ	95%	77.78%	42.42%	54.90%
単語＋隣接単語 2-gram	95%	72.22%	34.67%	46.85%
単語＋共起自立語 2-gram	95%	61.11%	41.51%	49.44%
単語＋限定共起 2-gram	95%	72.22%	40.00%	51.49%

限定共起語は正規表現集にある単語の中から単語リストが作られており，キーフレーズは，最初に正規表現で集めたツイート群から抽出しているため，本質的には正規表現で集めていることと変わらない．いずれも，学習データの特徴を文書 - 単語行列に反映させるために追加した素性ではあるが，その種となっているのは正規表現ルールである．ただし，限定共起語は特異値分解により人間が気付かなかった単語も素性に組み込むことが出来，キーフレーズに関しては，正規表現により集めたツイートの中から，人間が気付かずルールに書きこまなかったフレーズをも特徴量として素性に追加することが可能になった．これらのことから，この 2 つのみが F 値で単語 1-gram よりもよい結果をもたらしたと考えることができる．キーフレーズと個別正規表現は重なるものも多いが，キーフレーズの方が個別の正規表現を包含している関係にあるため，個別正規表現よりも再現率が高くなっていると解釈できる．

情報の方向性を担っている（「〜ている」のような現在形が，被害実態と未来の予測や注意喚起を分けている）ように見える「時制」などを扱うために，動詞文節を試したり，副詞や助動詞を含めた隣接バイグラムを扱うなどを試みてみたが，どのような素性がそのような効果を直接的に担っているかについては，結果を出すまでには至らなかった．

6 おわりに

震災対応に有用なツイートを抽出する情報抽出器を作成した．抽出器作成にあたり，震災時ツイートを初めて扱うことになった今回は，抽出対象の策定や特徴把握のため，震災時ツイートの全貌を明らかにする必要があったが，それを行う過程で効率的なクラスタリングを行うための手法と指標を提案し，それにより，震災時のツイートの俯瞰を得ることができた．機械学習によるターゲットツイートの複数種類同時抽出では，人手で作成した正規表現ルールを元に得られる素性を追加することの有効性が示唆されたものの，情報の方向を弁別できる決定的な素性またはその組み合わせの探査および検証は，引き続き今後の課題に据え置かれた．

一方，従来から次元圧縮や類義語処理，文書分類の観点で有効性が指摘されていた文書 - 単語行列の特異値分解は，特異値による重みづけを行うことで，さらに効果的な利用が行えることが示された．特異値分解後の重みづけの効果が最も顕著であったのは，クラスタリングにおけるチェイニング現象の緩和であった．このとき，特異値による重みづけは，クラスタリングの従来指標であるエントロピーや純度を必ず改善するわけではないが，自動クラスタリングの後に人手による修正作業が行われる際には重要な役割を果たすことがわかった．また，本研究において提案した，人手によるクラスタリング修正作業の負担を考慮した評価指標 (Easiness) についても，評価実験において実作業に要した時間との比較からその妥当性が示された．しかし特異値分解または特異値分解後に重みづけをすることが有効なのは，それらの行列に適切なカットオフを施した場合のみであり，適切な値の範囲はそれほど広くない．どのような値が適切であるかを定性的に説明するのは今後の課題である．

謝　辞

本稿執筆の機会を与えて下さった NHK 放送技術研究所の柴田部長および田中主任研究員，評価実験を手伝っていただいた同僚の木下奈々恵氏に心より感謝申し上げます．また，大変に有意義なコメントをいただきました査読者の方々，および何往復もやりとりいただいた事務局の方々にも同様に深く感謝いたします．

参考文献

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., and Sorensen, D. (1999). “LAPACK Users’ Guide.” 3rd edition. SIAM.

Berry, M. W. ed. (2004) “Survey of Text Mining.” Springer.

Berry, M. W. ed. (2008) “Survey of Text Mining II.” Springer.

Blei, D. M. and Ng, A. Y. (2003) “Latent Dirichlet Allocation.” *Journal of Machine Learning Research*, **3**(Jan), pp. 993–1022.

Cortes, C. and Vapnik, V. (1995). “Support-Vector Networks.” *Machine Learning*, **20**(3), pp. 273–297.

Crammer, K. and Singer, Y. (2000). “On the Learnability and Design of Output Codes for Multiclass Problems.” *Computational Learning Theory*, pp. 35–46.

Deerwester, S., Dumais, S., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). “Indexing by Latent Semantic Analysis.” *Journal of the American Society for Information Science*,

41(6), pp. 391–407.

Dongarra, J. J., Bunch, J. R., Moler, C. B., and Stewart, G. W. (1978). LINPACK Users Guide, Philadelphia: SIAM publications.

遠藤薫 (2012). メディアは大震災・原発事故をどう語ったか. 東京電機大学出版局.

Fisher, R. A., Sc. D., and F. R. S. (1936). "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics*, 7, pp. 179–188.

福田充 (2012). 大震災とメディア. 北樹出版.

稲泉連 (2012). IBC ラジオの 108 時間. 荒蝦夷（編）. その時，ラジオだけが聴こえていた. 竹書房.

Hofmann, T. (1999) "Probabilistic Latent Semantic indexing." In Proceedings of the 22th Annual International on SIGIR Conference on Research and development in information retrieval (SIGIR-99), pp. 50–57.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). "Data Clustering: A Review," *Journal of ACM Computing Surveys* (CSUR), **31** (3), pp. 264–323.

Karatzoglou, A., Smola, A., Hornik, K. and Achim, Z. (2004). "kernlab—An S4 Package for Kernel Methods in R." *Journal of Statistical Software*, **11**(9), pp. 1–20.

Karatzoglou, A., Meyer, D., and Hornik, K. (2006). "Support Vector Machines in R." *Journal of Statistical Software*, **15**(9), pp. 1–28.

片瀬京子 (2012). ラジオ福島の 300 日. 毎日新聞社.

Kobayashi, M. and Aono, M. (2004). "Vector Space Models for Search and Cluster Mining." Chapter 5 in book "Survey of Text Mining", Springer.

小林啓倫 (2011). 災害とソーシャルメディア. マイコミ新書.

Lin, H. T., Lin, C. J., and Weng, R. C. (2007) "A Note on Platt's Probabilistic Outputs for Support Vector Machines." *Machine Learning*, **68**(3), pp. 267–276.

Manning, C. D. and Schütze, H. (1999). "Foundations of Statistical Natural Language Processing." Section 5.3.1 The t test of Hypothesis Tesing in Collocations, The MIT Press Cambridge, Massachusetts London, England.

Manning, C. D., Raghavan, P., and Schutz, H. (2008). *Introduction to Information Retrieval*, Cambridge Univ. Press.

Neubig, G., 森 信介 (2012). 能動学習による効率的な情報フィルタリング. 言語処理学会第 18 回年次大会発表論文集, pp. 887–890.

岡崎直観, 成澤克麻, 乾健太郎 (2012). Web 文書からの人の安全・危険に関わる情報の抽出. 言語処理学会第 18 回年次大会発表論文集, pp. 895–898.

Platt, J. C. (2000). "Probabilistic Outputs for Support Vector Machines and Comparison to

Regularized Likelihood Methods.” In Smola, A., Bartlett, P., Schölkopf, B., and Schuurmans, D. eds. “Advances in Large Margin Classifiers,” MIT Press, Cambridge, MA.

Salton, G., Wong, A., and Yang, C. S. (1975). “A Vector Space Model for Automatic Indexing.” *Communications of the ACM*, **18**(11) pp. 613–620.

Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (1999). “Estimating the Support of a High-Dimensonal Distribution.” `http://research.microsoft.com/research/pubs/view.aspx?msr_tr_id=MSR-TR-99-87`.

立入勝義 (2011). 検証 東日本大震災そのときソーシャルメディアは何を伝えたか？ ディスカヴァー携書 066.

Tax, D. M. J. and Duin, R. P. W. (1999). “Support Vector Domain Description.” *Pattern Recognition Letters*, 20, pp. 1191–1199.

徳田雄洋 (2011). 震災と情報. 岩波新書.

Witten I. H., Paynter G. W., Frank E., Gutwin C. and Nveill-Manning C. G. (2005). “Kea: Practical automatic keyphrase extraction”. In Y.-L Theng and S. Foo, editors, Design and Usability of Digital Libraries: Case Studies in the Asia Pacific, pp. 129–152, Information Science Publishing, London.

略歴

平野真理子：2002 年慶應義塾大学大学院理工学研究科前期博士課程修了（工学修士）．在学時の研究テーマは核融合プラズマのシミュレーション．卒業後，アンケートや世論調査の設計および分析補助業務を経て，「顧客の声」等の自由記述文を定量的に扱い可視化する業務（メインは 100%人手による分類作業）に従事．現在は，分析の各種作業工程をより多く自動化することを目指し，研究を行っている．時折ソプラノ歌手として活動（東京二期会準会員）．

小早川　健：1993 年東北大学理学部物理学科卒業．1995 年東京大学理学系研究科物理学専攻修了．同年 4 月から日本アイ・ビー・エム株式会社．1999 年から日本放送協会．音声認識の研究を経て，現在は評判分析の研究に従事．

（2012 年 12 月 14 日　受付）
（2013 年 2 月 28 日　再受付）
（2013 年 3 月 29 日　採録）